

Designing a derivational resource for non-concatenative Morphology: the Hebrewnette database

Lior Laks

Bar-Ilan University

Ramat-Gan, Israel

Lior.Laks@biu.ac.il

Fiammetta Namer

UMR 7118 ATILF & Université de Lorraine

Nancy, France

fiammetta.namer@univ-lorraine.fr

Abstract

This paper presents a Derivational Database of Modern Hebrew (and more generally of Semitic languages) called Hebrewnette. The methodology adopted is based on adjusting the structure and properties of a database developed for the description of the derived lexicon of a Romance language (Démonette), and completing it to account for the specificities of the morphology of Semitic languages. We present the properties of Hebrewnette and the type of information it consists of, with special emphasis on both structural and semantic relations between words. Through a case study, we show how the annotations that are used allow us to verify theoretical hypotheses about non-concatenative morphology. The design of Démonette's annotation system makes its features, initially designed for French, suitable for capturing both morphological and semantic relations between Hebrew words, regardless of the type of morphology (concatenative or non-concatenative).

1 Introduction

This study presents the methodology of a derivational database of Hebrew (and more generally of Semitic languages) called Hebrewnette. The methodology adopted consists in adapting the structure of Démonette (Hathout and Namer, 2016; Namer and Hathout, 2020) a database developed for the description of the derived lexicon of French, and completing it to account for the specificities of the morphology of Semitic languages. Through a case study, we show how the annotations used allow us to verify theoretical hypotheses about non-concatenative morphology. The design of Hebrewnette relies on a word-based approach to morphology, whereas the tradition in the creation of tools (Daya et al., 2008) and lexical resources for Semitic languages (Neme, 2011)¹ is rather root-based (for an overview of theoretical approaches to Semitic morphology, see (Bat-El, 2017; Goldenberg, 1994; Ussishkin, 2005; Aronoff, 2007; Ravid, 2008; Berman, 2012; Faust, 2015; Kastner, 2020)). Hebrewnette provides a description of the derivational relations between Hebrew words in contrast to other types of database that relate mainly to inflectional paradigms. Finally, some works, even recent ones, point out the scarcity of freely available resources in Semitic languages, eg. in Arabic (El Haj et al., 2015). Hebrewnette (which is currently in a prototype stage) will contribute to fill this gap. Démonette, on which Hebrewnette is based, has been designed and implemented to represent the derivational relations within the French lexicon. Its realization is based on the following principles: (i) each entry is the relationship between two members of a derivational family; (ii) the same word participates in more than one entry; (iii) beside the classical base-to-derivative relations, entries in the database may correspond to cross-formations, or express a broader ancestor-descendant relation; (iv) both the words and their relation are identified by a set of morphological, phonological and semantic features.

2 Hebrew Morphology

Hebrew word formation relies highly on non-concatenative morphology, i.e. the formation via root and pattern (Berman, 1978; Bolozky, 1978; Schwarzwald, 1981; Ravid, 1990; Aronoff, 1994). The pattern

¹see also: <https://www.pealim.com/> for Modern Hebrew

indicates the prosodic structure of the word and it consists of the following elements: (i) consonantal slots; (ii) vocalic pattern; and in some cases (iii) affixes (Bat-El, 1994, 2017). For example, the verbs *diber* ‘speak_V’ and *tipes* ‘climb_V’ are formed in the CiCeC pattern. They share the vocalic pattern i-e and differentiate in their roots, d.b.r and t.p.s respectively. The verbs *hitraxec* ‘wash oneself_V’ and *hitragel* ‘get used to_V’ are formed in the hitCaCeC pattern, which consists of the prefix *hit-*, in addition to the vocalic pattern a-e. Words that share the same consonantal root typically share some semantic relations with different degrees of transparency, for example *hidpis* (hiCCiC) ‘print_V’, *hudpas* (huCCaC) ‘be printed_V’, *madpeset* (maCCeCet) ‘printer_N’ and *tadpis* (taCCiC) ‘printout_N’. Hebrew verbal patterns typically differ from each other with respect to transitivity and the semantic types of verbs that they host (see (Berman, 1978; Bolozky, 1978; Borer, 1991; Aronoff, 1994; Doron, 2003; Schwarzwald, 2008) and references therein). For example, CiCeC typically hosts active transitive verbs, e.g. *kivec* ‘shrink’, *nigev* ‘wipe’ and *xibek* ‘hug’, while hitCaCeC typically hosts intransitive verbs like inchoatives (*hitkavec* ‘become shrunk’), reflexives (*hitnagev* ‘wipe oneself’) and *hitxabek* ‘hug each other’). However, these only represent tendencies and there is no one-to-one correspondance between form and meaning of the patterns. For example, *hitʔalel* ‘abuse’ is formed in hitCaCeC but does not belong to any of the above mentioned semantic classes.

Within verb formation, non-concatenative formation is obligatory and every verb that enters the language must conform to one of the existing patterns. In contrast, the formation of nouns and adjectives is based on a variety of word formation strategies. Nouns, for example, can be raw (*cav* ‘turtle’), borrowed (*krason* ‘croissant’), and can be formed in both patterns and by affixation. For example, agent nouns are formed in the CaCaC pattern (*cayar* ‘painter’, *nagar* ‘carpenter’) and by affixation (*yam* ‘sea’ - *yamay* ‘sailor’).

3 From Démonette to Hebrewnette: overview

The founding principles of Démonette (Hathout and Namer, 2016; Namer and Hathout, 2020) that have been applied to Hebrewnette are the following:

- Each entry describes a derivational relationship between two lexemes.
- The entries form derivational families represented by connected graphs.
- A derivational relation regards any pair of members of the same family: it can connect an ancestor to a descendant (e.g. a derivative: *dansable_A* ‘danceable’ and its base: *danser_V* ‘dance’) or two derivatives of the same base (e.g. *danseur_{N_m}* ‘male dancer’ and *danseuse_{N_f}* ‘female dancer’), or two more distant elements of the family (e.g. *danser_V* and *indansable_A* ‘undanceable’). Each relation is coded according to its orientation (does it connect a derivative to its base? Two words derived from the same base? etc.) and complexity (i.e. the number of derivational steps required to connect the two words).
- The base is deliberately highly redundant: each lexical unit has as many derivational descriptions as it has connections within its family.
- In addition to the properties of its relation with other words, each lexical unit is defined by features independent of the relations in which it is found (e.g. its inflectional paradigm, part of speech, ontological category, frequency...)
- The (lexical and relational) properties are grouped into patterns that generalize the different levels of regularities that can be found in the constructed lexicon: phonological, semantic, morphological.

Like Démonette, Hebrewnette is represented in a tabulated format. Each entry is a pair of (noninflected) words (W_1 , W_2) belonging to the same derivational family. The morphological properties are divided between descriptions of the relations and descriptions of the words involved in these relations. The excerpt in Tab.1 gives an overview of the general organization of a Hebrewnette entry according to its

different properties. They are detailed in the following sections, in particular the features necessary for the expression of the non-concatenative morphology within Semitic languages.

As shown on the left part of Tab.1, each word is identified by its graphic form and phonetic transcription, its part of speech, and its English gloss. Formally, it is described by the pattern it belongs to, its root (and the type of root) and its vocalic structure, that is, its morphological structure (see §.4.2). When relevant, a feature encodes the variation between the morphological structure of a word and that of its pattern: for instance, the fact that the vowel /e/ in the noun *lemida* ‘learning’ is not predicted by its pattern CCiCa (see details in §.4.1 and Tab.2.4, column P_i to W_i).

Finally, each word is annotated by means of its ontological properties (Semantic Type, Semantic Subtype) and its argument structure (features Transitivity and Argument Structure). In Tab.1, the value ‘dyn’ of Semantic Type for *lamad* and *limes* indicates that both verbs are dynamic predicates. *lamad* is a regular active transitive predicate (Semantic Subtype= act, Transitivity=trans.), which is reflected by the value XY of its Argument Structure (someone_X studies something_Y). The Semantic Subtype and Transitivity features of *limes* are valued causative and transitive, respectively, because *limes* introduces a causative argument W in its argument structure, with respect to *lamad* argument structure (someone_W teaches someone else_X something_Y).

The relation between two words is described according three dimensions, for reasons explained in §.4.3 (see right part of Tab.1):

- The orientation and complexity of the relation (is W₁ derived from W₂, W₂ from W₁ ? none of them is derived from the other? How many derivational steps are there between W₁ and W₂?) are examined separately from the structural and semantic points of view, see also Tab.4;
- The phonological dimension of the relation concerns the possible variation between the two words, and/or between their roots, see also Tab.3;
- The semantic relation is paraphrased by a gloss that cross-defines W₁ and W₂. Here, the cross-definition of *lamad* and *limes* illustrates the causative relation between the two verbs and between their arguments.

	Word ₁	Word ₂	Relation between Word ₁ and Word ₂	
Written form	למד	לימד	Formal orientation	NA
Phon transc	<i>lamad</i>	<i>limes</i>	Formal complexity	simple
Transl	study	teach	W ₁ /W ₂ Phon alternation	NA
PoS	v	v	Relation bwn roots	=
Pattern	CaCaC	CiCeC	Semantic orientation	W ₁ → W ₂
Root	l.m.d	l.m.d	W ₁ /W ₂ cross-definition	“when W limes X Y, then X lamad Y”
Root type	regular	regular		
Morphological representation	[aa]	[ie]		
Pattern-to-Word phon. altern.	NA	NA		
Semantic type	dyn	dyn		
Semantic subtype	act	caus		
Transitivity	trans.	trans.		
Argument structure	XY	WXY		

Table 1: The Hebrewnet database: an excerpt

4 The Hebrewnet database

In the following, we provide some examples of information needed to accurately represent the properties of words constructed by non-concatenative morphology. These features serve various purposes: represent each derivational relation and each word involved in it in terms of roots and patterns (§.4.1) as well as the (relation between) root classes (§.4.2), and describe meaning-form asymmetry between the formal and the semantic orientations of the derivational relation (§.4.3).

4.1 Roots, patterns, affixes and structural variations

As we have just seen, the representation of non-concatenative derivations involves different annotations illustrated in Table 2. Some features relate to the words involved in the relation: they are distinguished according to whether or not they have a pattern (columns **Pattern P₁** and **Pattern P₂**). When the word has no pattern, it may be ‘borr.(owed)’ (e.g. *spam* in T2.1) or ‘raw’ (e.g. *yam* and *yami* ‘of sea’, in T2.3). When relevant, the representation of the pattern is completed by the description of the root (e.g. l.m.d for column **R₂** in T2.4), and that of each word structure (at columns **W₁ Struct.** and **W₂ Struct.**). The structure of a word consists in a vocalic pattern (e.g. |oe| for *lomed* in T2.4), possibly completed by affixes belonging to the pattern (e.g. *ti* and *et* in *ti|0o|et*, in T2.2) and autonomous affixes (e.g. the suffix *-i* in *|iu|+i*, in T2.6). When relevant, the indication of a phonological shift between the representation of the word and that of its pattern is also provided. For instance, the annotation: $\emptyset/e_{W_2}^{V_1}$ in T2.4, column **P_i to W_i** indicates the insertion of the vowel /e/ in position V₁ of W₂, that is, between the first and the second consonants of the word root. On the CCiCa P₂ pattern, V₁ is empty (the absence of the vowel is represented by the value ‘0’) whereas it is filled with /e/ in the W₂ *lemida*. The CCiCa pattern typically has an initial consonantal cluster (CC) and vowel insertion does not occur, as illustrated with *šmira*, in T2.5. Other features are used to describe the structure of the relation itself (column **Structure of relation**), and the phonological variation between W₁ and W₂. For example, in T2.2, column **W₁/W₂ phono. alt.**, there is a /v/ to /b/ variation on the consonant position C₂, between *gavar* and *tigboret*.

	W ₁	W ₂	Pattern P ₁	Pattern P ₂	R ₁	W ₁ struct.	R ₂	W ₂ struct.	P _i to W _i	Structure of the relation	W ₁ /W ₂ phono. alt.
T2.1	<i>spam</i> ‘spam _N ’	<i>hispi</i> ‘spam _V ’	borr.	hiCCiC		0a	s.p.m	hi 0i		CCaC/P2	
T2.2	<i>gavar</i> ‘increase _V ’	<i>tigboret</i> ‘reinforcement _N ’	CaCaC	tiCCoCet	g.b.r	laa	g.b.r	ti 0o et		P1/P2	v/b ^{C2}
T2.3	<i>yam</i> ‘sea _N ’	<i>yami</i> ‘marine _A ’	raw	raw		W		W+i		W/W+i	
T2.4	<i>lomed</i> ‘learner _N ’	<i>lemida</i> ‘learning _N ’	CoCeC	CCiCa	l.m.d	oe	l.m.d	0i a	$\emptyset/e_{W_2}^{V_1}$	P1/P2	
T2.5	<i>šomer</i> ‘guard _N ’	<i>šmira</i> ‘guarding _N ’	CoCeC	CCiCa	š.m.r	oe	š.m.r	0i a		P1/P2	
T2.6	<i>limud</i> ‘teaching _N ’	<i>limudi</i> ‘educational _A ’	CiCuC	CiCuC+i	l.m.d	iu	l.m.d	iu +i		P1/P2	

Table 2: Formal representation of (relations between) words and patterns

4.2 Root types and inter-family relations

Words sharing the same root typically belong to the same morphological family (on the other hand, some families may consist of words without roots, as in T2.3). Morphological families form paradigms. The root description (Table 3) is information specific to each word. Roots are classified according to different types. They are regular (‘r’) if they contain three consonants (for example, d.r.x, in T3.2), ‘r-4’ if they are quadrilateral. In this case the dot ‘.’ is used to group clusters (as t.dr.x in T3.4). When a pattern surfaces as a wordform with only 2 consonants (for instance *rac* in T3.5) the historical value of the missing root consonant is noted in capitals (e.g. /w/ of r.W.c in T3.5). Other values, not illustrated here, complete this tagset: for example, they indicate when the same root (e.g. s.p.r) corresponds to disjoint families with homonyms (*siper* ‘tell_V’ vs. *siper* ‘cut hair_V’) or polysemes (*xafar* ‘dig_V’ vs. *xafar* ‘talk too much_V, drill one’s mind (metaphorically)_V’).

By default, a relation connects two items that share the same root (provided they belong to a pattern, compare T2.3 to T2.4). However, there are relations that connect items with different roots. These particular relations are characterized by adding a consonant to the root on word W₂, as in T3.3, where

	W ₁	W ₂	Patt. P ₁	Patt. P ₂	R ₁	R ₁ type	R ₂	R ₂ type	R ₁ to R ₂
T3.1a	<i>mahal</i> 'mix _V '	<i>tamhil</i> 'mix _N '	CaCaC	taCCiC	m.h.l	r	m.h.l	r	
T3.1b	<i>hidpis</i> 'print _V '	<i>tadpis</i> 'printout _N '	hiCCiC	taCCiC	d.p.s	r	d.p.s	r	
T3.2a	<i>hidrix</i> 'guide _V '	<i>tadrix</i> 'briefing _N '	hiCCiC	taCCiC	d.r.x	r	d.r.x	r	
T3.2b	<i>hidrix</i> 'guide _V '	<i>hadraxa</i> 'guidance _N '	hiCCiC	haCCaCa	d.r.x	r	d.r.x	r	
T3.2c	<i>tadrix</i> 'briefing _N '	<i>hadraxa</i> 'guidance _N '	taCCiC	haCCaCa	d.r.x	r	d.r.x	r	
T3.3	<i>tadrix</i> 'briefing _N '	<i>tidrex</i> 'debrief _V '	taCCiC	CiCeC	d.r.x	r	t.dr.x	r-4	CCC/tCCC
T3.4a	<i>tidrex</i> 'debrief _V '	<i>tidrux</i> 'debriefing _N '	CiCeC	CiCuC	t.dr.x	r-4	t.dr.x	r-4	
T3.4b	<i>tidrex</i> 'debrief _V '	<i>tudrax</i> 'be debriefed _V '	CiCeC	CuCaC	t.dr.x	r-4	t.dr.x	r-4	
T3.5	<i>rac</i> 'run _V '	<i>rica</i> 'running _N '	CaCaC	CCiCa	r.W.c	r _{C2=W}	r.W.c	r _{C2=W}	

Table 3: Root classification

d.r.x → t.dr.x. In that case, the variation between roots R₁ and R₂ is specified, for example, with CCC/tCCC. This type of relationship creates a new family, and its members share the new root. The two families form different paradigms. We can illustrate this observation with *tadrix* 'briefing_N'/ *tidrex* 'debrief_V' (T3.3):

- The taCCiC pattern, which includes the prefix *ta-*, is used for the formation of different kinds of nouns that can be related to verbs in different patterns, e.g. *mahal* 'mix_V (liquids)' - *tamhil* 'mix_N' (T3.1a), *hidpis* 'print_V' - *tadpis* 'printout_N' (T3.1b). The noun *tadrix* 'briefing' is formed in the taCCiC pattern, and is semantically related to the hiCCiC transitive verb *hidrix* 'guide_V' (T3.2a) and the haCCaCa action noun *hadraxa* 'guidance_N' (T3.2c). The three words are interconnected (T3.2a, 2b, 2c) and form a derivational family sharing the consonantal root d.r.x.
- As T3.3 shows, the verb *tidrex* 'debrief_V' is formed in the CiCeC pattern based on the noun *tadrix*, taking the t consonant of the derivational prefix *ta-* as part of the new root t.dr.x. The CiCeC pattern is paradigmatically connected to the CiCuC pattern of action noun (*tidrux* 'debriefing_N' in T3.4a) and to the verbal passive CuCaC pattern (*tudrax* 'be debriefed_V' in T3.4b).

We can see that the pattern CiCeC of W₂ (*tidrex*) induces new types of relations within its new family. These relations are paradigmatically determined. We can therefore say that a relation like *tadrix/tidrex* serves to connect two paradigms.

4.3 Meaning-form discrepancies: relations with diverging orientations

In Démonette, the value of the orientation feature indicates which of the two related words is the base (or the ancestor) of the other. Non-concatenative morphology is such that the formal orientation is often impossible to determine. For instance, in the *cilem/cilum* relation there is no formal clue to decide if *cilem* 'photograph_V' is the base of *cilum* 'photography_N' or is derived from it. By distinguishing semantic orientation and formal orientation these two aspects are dissociated. Therefore each derivational relation in a family can be properly described according to the value combination of these two independent features. Table 4 shows several cases of such combinations; orientations (columns 4 and 5) are symbolized by arrows, f₁ and f₂ stand for the form of W₁ and W₂ respectively, s₁ and s₂ represent their semantic content.

- base word → derived word *regular* orientation (T4.1): *macléma* 'camera_N' is more complex both formally and semantically than *cilem* 'photograph_V' (we assume that W₂ is semantically more complex than W₁ if the semantic content of W₂ includes at least one additional predicate or operator compared to W₁: here, W₂ denotes the instrument used to perform the action described by W₁).

- base word → derived word *semantic* orientation (T4.2): *šavir* ‘breakable_A’ is more complex than *šavar* ‘break_V’, whereas the formal orientation cannot be determined.
- base word → derived word *formal* orientation (T4.3): the structure of *hitkavec* ‘get shrunk_V’ is more complex than that of *kivec* ‘shrink_V’. On the other hand, no semantic orientation can be assigned to the relation: it is unclear whether the intransitive predicate is built from the transitive one, or vice-versa (Haspelmath, 1993).
- indirect semantic relation (T4.4): the formal orientation between *kuvac* ‘be shrunk_V’ and *kavic* ‘shrinkable_A’ is indeterminate, and the semantic content of the two words are defined based on a common morphosemantic base (*kivec* ‘shrink_V’)
- two more combinations are illustrated in T4.5. The semantic contents of the agent noun *calam* ‘photographer_N’ and the instrument noun *maclema* are not directly related to one another, but they are semantically linked to the common verb ancestor *cilem*, and *maclema* can be formally derived from *calam*.
- double indeterminacy: in T4.6 the noun *šuman* and the adjective *šamen* are of the same formal complexity and share the same semantic content (‘fat’).

	W ₁	W ₂	Struct. of the rel.	Form. orient.	Sem. orient.
T4.1	<i>cilem</i> ‘photograph _V ’	<i>maclema</i> ‘camera _N ’	CiCeC/maCCeCa	f ₁ → f ₂	s ₁ → s ₂
T4.2	<i>šavar</i> ‘break _V ’	<i>šavir</i> ‘breakable _A ’	CaCaC/CaCiC	–	s ₁ → s ₂
T4.3	<i>kivec</i> ‘shrink _V ’	<i>hitkavec</i> ‘become shrunk _V ’	CiCeC/hitCaCeC	f ₁ → f ₂	–
T4.4	<i>kuvac</i> ‘be shrunk _V ’	<i>kavic</i> ‘shrinkable _A ’	CuCaC/CaCiC	–	s ₁ ↔ s ₂
T4.5	<i>calam</i> ‘photographer _N ’	<i>maclema</i> ‘camera _N ’	CaCaC/maCCeCa	f ₁ → f ₂	s ₁ ↔ s ₂
T4.6	<i>šamen</i> ‘fat _A ’	<i>šuman</i> ‘fat _N ’	CaCeC/CuCaC	–	–

Table 4: Structural vs. semantic orientation of a relation

5 Case study: maCCuC formation

Some Hebrew adjectives have doublets that are formed in the maCCuC pattern, mostly in a jocular manner. The adjectives *maxrid* (1a)² and *maxrud* (1b)³, both denote ‘awful’, share the consonants x . r . d, but are formed in different patterns. A similar case is presented in (2) for *misken*⁴ and *maskun*⁵ ‘poor_A’.

- | | |
|--|---|
| <p>(1) a. lavašti jins maxrid
‘I wore an awful pair of jeans’</p> <p>b. hi xorešet al oto jins maxrud
‘she wears the same awful jeans’</p> | <p>(2) a. eyze misken, ma hu asa la
‘what a poor (guy), what did he do to her?’</p> <p>b. eyze maskun, kol paam ani yocet alexa
‘what a poor (guy), I lash out at you every time’</p> |
|--|---|

Not all speakers accept maCCuC forms like the ones in (1b) and (2b) (Bolozky, 1999, 2000), yet web searches reveal that they are productive. In contrast to cases like (1)-(2), there are many adjectives that do not have maCCuC counterparts, e.g. *metunaf* – **matnuf* ‘filthy’. Why is it so? maCCuC formation (and

²<https://bike.co.il/?p=2239>

³<http://tmi.maariv.co.il/style/Article-609396>

⁴https://www.tiktok.com/@einabl_253/video/6948081577649818881

⁵<https://www.inn.co.il/Forum/Forum.aspx/t851240>

lack thereof) can be predicted based on structural and semantic properties of the base adjective. From the semantic point of view, maCCuC adjectives must have negative meaning, and therefore adjective like *maksim* ‘charming’ and *meratek* ‘fascinating’ do not have such doublets (**maksum*, **martuk*). maCCuC adjectives can be derived from adjectives in different patterns that are not marked for specific semantic meaning, e.g. maCCiC, muCCaC. This derivation is not oriented formally because both patterns are equally complex as they both consist of a prefix. The derivation is semantically oriented from maCCiC or muCCaC to maCCuC because a negated property is semantically more complex than the corresponding unmarked one.

On the structural dimension, adjectives with maCCuC doublets must have medial consonant clusters. maCCuC formation is faithful to the base, as it involves vowel(s) changes and preserves the syllabic structure (T5-a). This brings about structural transparency between the forms. maCCuC formation based on adjectives without medial clusters involves more changes of the base, especially modification of the syllabic structure, and therefore it is highly rare (T5-b) or unattested (T5-c,d). Unattested forms are not included in Hebrewnet, we add them here just for the sake of demonstration.

This difference can be predicted from the string distance between the ‘regular’ form (W_1) and its doublet W_2 . The greater the difference, the higher the probability that W_2 is either very rare, or unattested. Distances can be computed by means of a string metric. In Table 5, we use a measure parametrized such that string modification is weighed according to the distance from the original syllabic structure. Therefore, vowel substitution is twice “cheaper” as prefix insertion or deletion. Moreover, it weights four times less than vowel deletion or insertion, because the latter transformation involves consonant (de)clusterization, that is, either breaking consonant clusters that exist in the base, or creating consonant clusters that are not part of the base. A maCCuC adjective occurs when the distance with respect to the ‘regular’ negative adjective is smaller than 4 or equals to it. Since Hebrewnet encodes both semantic and structural information of each word and the relations between words, this allows to predict which adjectives are more likely to have maCCuC doublets.

	W_1	W_2	W_1 Str.	W_2 Str.	W_1/W_2 string operations	W_1/W_2 string distance
a. Frequent maCCuC formations						
T5.a	<i>maxrid</i>	<i>maxrud</i> ‘awful _A ’	ma 0i	ma 0u	$V_{subs}: /i/ > /u/$	1
	<i>misken</i>	<i>maskun</i> ‘poor _A ’	mi 0e	ma 0u	$V_{subs}: /i/ > /a/; /e/ > /u/$	2
b. Unfrequent (T5.b) or unattested (T5.c,d) maCCuC formations						
T5.b	<i>metoraf</i>	<i>matruf</i> ‘crazy _A ’	me ua	ma 0u	$V_{subs}: /e/ > /a/; /a/ > /u/$ $V_{del}: /u/ > 0$	6
T5.c	<i>metunaf</i>	<i>*matnuf</i> ‘filthy _A ’	me ua	ma 0u	$V_{subs}: /e/ > /a/; /a/ > /u/$ $V_{del}: /u/ > 0$	6
T5.d	<i>satum</i>	<i>*mastum</i> ‘blockheaded _A ’	au	ma 0u	Prefix: <i>ma-</i> ; $V_{del}: /a/ > 0$	6

Table 5: Likelihood of maCCuC adjective doublets formation

Interestingly, adjectives with negative meaning without maCCuC counterparts have semi-counterparts in Segolate (Bat-El, 2012; Shany-Klein and Ornan, 1992) patterns like CeCeC or CaCeC. We relate to them as semi-counterparts or “semi-doublets” because unlike maCCuC, which is used for the formation of adjectives, these Segolate patterns usually serve for the formation of nouns, e.g. *satum* ‘thickheaded’, *setem* ‘a thickheaded person’, *metunaf* ‘filthy’, *tanev* ‘a filthy person’. These segolate forms have peculiar behavior as they are not inflected for gender and number, unlike Hebrew animate nouns. For example, *metunaf* modifies only masculine nouns and its feminine form is *metunef-et*, while *tanev* relates to both genders. Regardless of the special status of these Segolate forms, they tend to be in complementary distribution with maCCuC forms with respect to marking the negative meaning of existing adjectives.

Similarly to the case of maCCuC doublet formation, the formation of CeCeC or CaCeC forms does not modify the syllabic structure of the base. Both types of formation involve faithfulness to the base.

Examine again the attested adjective *metunaf*. It doesn't have a maCCuC counterpart (**matnuf*, T5-c) because such formation would involve vowel deletion (in addition to vowel substitution), which creates a consonant cluster and therefore infringes syllabic structural faithfulness with respect to *metunaf*: this results in a distance of 6 between the two forms. In contrast, the formation of *tanef* (T6-a) is less pricy, because its distance from *metunaf* is only 4: the prefix *ma-* is deleted and vowels are substituted, but the syllabic structure of the two stems is the same. There are some cases in which the Segolate pattern formation is even less pricy, e.g. *satum* – *setem* (T6-b) where the two stems share the same syllabic structure. In both cases in Table 6, there is no modification of the syllabic structure of the base and therefore the formation of Segolate forms is cheaper than maCCuC forms.

Existing adjectives with a medial consonant cluster do not have Segolate counterparts for the same reason, namely such formation would change the syllabic structure of the base by breaking a consonant cluster, in addition to other changes. The adjective *maxrid* 'awful', for example, does not have a Segolate semi-counterpart like **xered* (T6-c) because this relation would imply vowel insertion that breaks the *xr* cluster, deletion of the prefix *ma-* and vowel substitution, corresponding to a distance of 7 between them.

	W ₁	W ₂	W ₁ Str.	W ₂ Str.	W ₁ /W ₂ string operations	W ₁ /W ₂ string distance
T6.a	<i>metunaf</i>	<i>tanef</i>	me ua	ae	V _{subs} : /e/ > /a/; /a/ > /u/ Prefix del.: <i>me-</i>	4
T6.b	<i>satum</i>	<i>setem</i>	au	ee	V _{subs} : /a/ > /e/; /a/ > /u/	2
T6.c	<i>maxrid</i>	<i>*xered</i>	ma 0i	ee	Prefix del.: <i>ma-</i> V _{subs} : /i/ > /e/; V _{ins} : /0/ > /e/	7

Table 6: CeCeC and CaCeC doublets formation

6 The Hebrewnette prototype

Hebrewnette is a prototype of 250 entries. The description of each entry is the product of 37 features. The Hebrewnette core is made up of 160 entries, corresponding to 127 lexemes and 19 families. They have been encoded to test the robustness of the database. These entries combine one or several of the characteristics specific to Hebrew derivation that we have presented in this article: mismatched formal and semantic orientations, non-triconsonant roots, absence of pattern, phonological alternations, structural variations, etc. The 10 other derivational families included in the current version of Hebrewnette have been generated and annotated semi-automatically. Based on an initial list of 10 CiCeC verbs, we relied on the nature fundamentally paradigmatic of the Hebrew verbal lexicon to implement the following predictions:

- CiCeC verbs are likely to realize active, transitive, dynamic predicate, e.g. *xibek* 'hug_V', *kivec* 'shrink_V', *nihel* 'manage_V';
- they are related to a CiCuC action noun (*xibuk* 'hug_N', *nihul* 'management_N'), a resultative adjective in the meCuCaC participle pattern (*menohal*⁶ 'managed_A'). CiCeC is also derivationally related to the meCaCeC participle pattern that can surface as an adjective (*mexabek* 'hugging_A'), an agent noun (*menahel* 'manager_N') or an instrument noun;
- when attested, their hitCaCeC related verb is intransitive, typically inchoative (*hitkavec* 'become shrunk_V'), reflexive (*hitraxec* 'wash oneself_V') or reciprocal (*hitxabek*, 'hug each other_V').

⁶The /u/ to /o/ variation between the pattern meCuCaC and the word *menohal* is due to the fact that the second consonant of the root /h/ is a glottal stop.

From these 10 CiCeC verbs, the program produced 70 new annotated lexemes (after manual verification, 20 of them are discarded): each CiCeC verb is the source of a family of 6 members on average. Insofar as each member in a family is linked to all the others, this amounts to supplement the 160 initial wordpairs with 90 new fully documented entries.

7 Conclusions

This paper presented the main principles of designing Hebrewnette, a derivational database for Hebrew, and its properties. We accounted for the adaptations that were made on the Démonette database, which was originally designed for Romance Morphology. Focus was on non-concatenative formation, which is highly typical of Hebrew and Semitic languages in general. We outlined the way words were coded with respect to their root and pattern. Taking a word-based approach for word formation, Hebrewnette is also based on coding relations between words, and specifically for Hebrew, relations between roots and patterns. It is based on separate description of semantic and structural relations so that each type of relation can be examined according to different criteria, e.g. direction of derivation (if any). We examined a case study of doublet formation of adjectives in the Hebrew maCCuC pattern, and showed that the way words and their relations are coded in Hebrewnette can account for the likelihood of such doublet formation. While such doublet formation is semantically motivated in order to mark adjectives as carrying negative meaning, the likelihood of doublet formation is based on structural relations between the existing adjective and its doublet and the degree of faithfulness between them, namely the types of changes that the doublet formation requires. We showed that the proposed design of Hebrewnette allows the representation of the role of faithfulness in word formation.

The features and feature values in the Hebrewnette database intertwine with the content of Démonette, to account for the particularities of languages with non-concatenative morphology. However these additions do not compromise the architecture of Démonette, the global structures of the two databases are superimposable, which allows us to envisage a total interoperability between the two systems (and more widely between the morphologies of Romance languages and Semitic languages). We have shown that the combinability of features allows us to empirically verify hypotheses, which confirm the validity of Word-based approaches in non-concatenative morphology. Nonetheless, just like the Démonette database from which it is inspired, Hebrewnette allows for a multi-theoretical consultation / analysis of derivational relations, in the sense that it gives access not only to word-and-pattern relations (in order to be suited to the family and paradigms principles of derivation), but also to roots and root-and-pattern relations (in accordance with the needs of the root-based approaches to Semitic morphology).

Acknowledgments

This research has been realized as part of the project Demonext⁷, supported by the Agence Nationale de la Recherche, grant number: ANR-17-CE23-0005. It has been also been supported by the Chateaubriand Fellowship Program of the French Embassy in Israel.

References

- Mark Aronoff. 1994. *Morphology by Itself. Stem and Inflectional Classes*. MIT Press, Cambridge, MA.
- Mark Aronoff. 2007. In the Beginning was the word. *Language* 83:803–830.
- Outi Bat-El. 1994. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory* 12:572–596.
- Outi Bat-El. 2012. Prosodic alternations in Modern Hebrew segolates. In Malka Muchnik and Zvi Sadan, editors, *Studies on Modern Hebrew and Jewish Languages*, Carmel Press, Jerusalem, pages 116–129.
- Outi Bat-El. 2017. Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In Claire Bowern, Laurence Horn, and Raffaella Zanuttini, editors, *On Looking into Words (and beyond)*, Language Science Press, Berlin, pages 115–135.

⁷<https://www.demonext.xyz/>

- Ruth Berman. 1978. Modern Hebrew structure. Report, University Publishing Projects.
- Ruth Berman. 2012. Revisiting roots in Hebrew: A multi-faceted view. In Malka Muchnik and Zvi Sadan, editors, *Studies on Modern Hebrew and Jewish Languages*, Carmel Press, Jerusalem, pages 132–154.
- Shmuel Bolozky. 1978. Word formation strategies in MH verb system: denominative verbs. *Afroasiatic Linguistics* 5:1–26.
- Shmuel Bolozky. 1999. *Measuring productivity in word formation: the case of Israeli Hebrew*. Brill, Leiden.
- Shmuel Bolozky. 2000. Stress placement as a morphological and semantic marker in Israeli Hebrew. *Hebrew Studies* 41:53–82.
- Hagit Borer. 1991. The causative-inchoative alternation: a case study in parallel morphology. *The Linguistic Review* 8:119–158.
- Ezra Daya, Dan Roth, and Shuly Wintner. 2008. Identifying semitic roots: Machine learning with linguistic constraints. *Computational Linguistics* 34(3):429–448.
- Edit Doron. 2003. Agency and voice: The semantics of the Semitic templates. *Natural Language Semantics* 11:1–67.
- Mahmoud El Haj, Udo Kruschwitz, and Chris Fox. 2015. [Creating language resources for under-resourced languages: methodologies, and experiments with Arabic](https://doi.org/https://doi.org/10.1007/s10579-014-9274-3). *Language Resources and Evaluation* 49:549–580. <https://doi.org/https://doi.org/10.1007/s10579-014-9274-3>.
- Noam Faust. 2015. A novel, combined approach to semitic word-formation. *Journal of Semitic Studies* LX(2):287–316.
- Gideon Goldenberg. 1994. Principles of Semitic word-structure. In Gideon Goldenberg and Schlomo Raz, editors, *Semitic and Cushitic Studies*, Harrassowitz Verlag, Wiesbaden, pages 10–45.
- Martin Haspelmath. 1993. *More on the typology of inchoative / causative verb alternations*, John Benjamins, Amsterdam/Philadelphia, pages 87–111.
- Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Itamar Kastner. 2020. *Voice at the interfaces: The syntax, semantics and morphology of the Hebrew verb*. Language Science Press, Berlin. <https://doi.org/10.5281/zenodo.3865067>.
- Fiammetta Namer and Nabil Hathout. 2020. [Paradis and démonette –from theory to resources for derivational paradigms](https://doi.org/10.14712/00326585.001). *The Prague Bulletin of Mathematical Linguistics* 114:5–33. <https://doi.org/10.14712/00326585.001>.
- Alexis Amid Neme. 2011. [A lexicon of Arabic verbs constructed on the basis of semitic taxonomy and using finite-state transducers](https://halshs.archives-ouvertes.fr/halshs-01186723). In Benoît Sagot, editor, *WoLeR 2011*. pages 79–86. [halshs-01186723](https://halshs.archives-ouvertes.fr/halshs-01186723).
- Dorit Ravid. 1990. Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica* 24:289–347.
- Dorit Ravid. 2008. Parsimony and efficacy: The dual binyan system of hebrew. In *13th International Morphology Meeting (13th IMM)*.
- Ora R. Schwarzwald. 1981. *Grammar and Reality in the Hebrew Verb*. Bar Ilan University Press, Ramat Gan.
- Ora R. Schwarzwald. 2008. The special status of nif'al in hebrew. In Sharon Armon-Lotem, Gabi Danon, and Susan Rothstein, editors, *Current Issues in Generative Hebrew Linguistics*, John Benjamins, Amsterdam, pages 61–75.
- Michal Shany-Klein and Uzzi Ornan. 1992. Analysis and generation of hebrew segolate nouns. In Uzzi Ornan, Gideon Arieli, and Edit Doron, editors, *Hebrew Computational Linguistics*, Ministry of Science and Technology, Jerusalem, pages 39–51.
- Adam Ussishkin. 2005. A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory* 23:169–218.